Recommendations for a step-wise comparative approach to the evaluation of new screening tests for colorectal cancer

Young, Graeme P.; Senore, Carlo; Mandel, Jack S.; Allison, James E.; Atkin, Wendy S.; Benamouzig, Robert

Published in:
Cancer

DOI:
10.1002/cncr.29865

Publication date:
2016

Commentary

Recommendations for a Step-Wise Comparative Approach to the Evaluation of New Screening Tests for Colorectal Cancer

Graeme P. Young, MD, FRACP, FTSE, AGAF[1]; Carlo Senore, MD, MSc[2]; Jack S. Mandel, PhD, MPH[3]; James E. Allison, MD, FACP, AGAF[4]; Wendy S. Atkin, MPH, PhD[5]; Robert Benamouzig, MD, PhD[6]; Patrick M. M. Bossuyt, PhD[7]; Mahinda De Silva, MB, BS, FRACP[8]; Lydia Guittet, MD, PhD[9]; Stephen P. Halloran, MBE, FRCPath[10]; Ulrike Haug, PhD[11]; Geir Hoff, MB, ChB, PhD[12]; Steven H. Itzkowitz, MD, FACP, FACG, AGAF[13]; Marcis Leja, MD, MBA, PhD, AGAF[14]; Bernard Levin, MB, BCh, FACP[15]; Gerrit A. Meijer, MD, PhD[16]; Colm A. O'Morain, MD[17]; Susan Parry, MbCHB, FRACP18[18]; Linda Rabeneck, MD, MPH, FRCPC[19]; Paul Rozen, MD[20†]; Hiroshi Saito, MD, PhD[21]; Robert E. Schoen, MD, MPH[22]; Helen E. Seaman, BSc, PhD[23]; Robert J. C. Steele, MD, FRCS[24]; Joseph J. Y. Sung, MD, PhD[25]; and Sidney J. Winawer, MD[26]

BACKGROUND: New screening tests for colorectal cancer continue to emerge, but the evidence needed to justify their adoption in screening programs remains uncertain. **METHODS:** A review of the literature and a consensus approach by experts was undertaken to provide practical guidance on how to compare new screening tests with proven screening tests. **RESULTS:** Findings and recommendations from the review included the following: Adoption of a new screening test requires evidence of effectiveness relative to a proven comparator test. Clinical accuracy supported by programmatic population evaluation in the screening context on an intention-to-screen basis, including acceptability, is essential. Cancer-specific mortality is not essential as an endpoint provided that the mortality benefit of the comparator has been demonstrated and that the biologic basis of detection is similar. Effectiveness of the guaiac-based fecal occult blood test provides the *minimum* standard to be achieved by a new test. A 4-phase evaluation is recommended. An initial retrospective evaluation in cancer cases and controls (Phase 1) is followed by a prospective evaluation of performance across the continuum of neoplastic lesions (Phase 2). Phase 3 follows the demonstration of adequate accuracy in these 2 prescreening phases and addresses programmatic outcomes at 1 screening round on an intention-to-screen basis. Phase 4 involves more comprehensive evaluation of ongoing screening over multiple rounds. Key information is provided from the following parameters: the test positivity rate in a screening population, the true-positive and false-positive rates, and the number needed to colonoscope to detect a target lesion. **CONCLUSIONS:** New screening tests can be evaluated efficiently by this stepwise comparative approach. ***Cancer* 2016;122:826-39**.

KEYWORDS: colonoscopy, colorectal cancer, fecal occult blood test, molecular diagnostics, screening test.

Corresponding author: Graeme P. Young, MD, FRACP, Flinders Center for Innovation in Cancer, Flinders University, Level 3, Flinders Medical Center, Bedford Park, Adelaide, SA 5042, Australia; Fax: (011) 61-8-8204-3943; graeme.young@flinders.edu.au

[1]Flinders Center for Innovation in Cancer, Flinders University, Adelaide, South Australia, Australia; [2]Reference Center for Epidemiology and Cancer Prevention, Piedmont Regional Center for Preventive Oncology, City Health and Science University Hospital of Turin, Turin, Italy; [3]Environmental and Occupational Medicine, University of Minnesota, Minneapolis, Minnesota; [4]Division of Gastroenterology, University of California, San Francisco and Kaiser Division of Research, Oakland, California; [5]Gastrointestinal Epidemiology, Imperial College, London, United Kingdom; [6]Gastroenterology Department, Avicenne Hospital, Paris 13 University, Paris, France; [7]Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands; [8]Department of Gastroenterology, Repatriation General Hospital, Adelaide, South Australia, Australia; [9]Unit 1086, French National Institute for Health and Medical Research, Cancers and Preventions Center, Caen University Hospital, Caen, France; [10]Faculty of Health and Medical Sciences, University of Surrey, Guildford, United Kingdom; [11]Department of Clinical Epidemiology, Leibniz Institute for Prevention Research and Epidemiology, Bremen, Germany; [12]Telemark Hospital, Skein Cancer Registry of Norway, University of Oslo, Oslo, Norway; [13]Gastrointestinal Division, Icahn School of Medicine at Mount Sinai, New York, New York; [14]Digestive Diseases Center, GASTRO, Faculty of Medicine, University of Latvia, Riga, Latvia; [15]Division of Cancer Prevention, The University of Texas MD Anderson Cancer Center, Houston, Texas; [16]Netherlands Cancer Institute, Amsterdam, The Netherlands; [17]Faculty of Health Science, Trinity College Dublin, Dublin, Ireland; [18]Ministry of Health Bowel Cancer Program, Auckland Hospital, Auckland, New Zealand; [19]Prevention and Cancer Control, Cancer Care Ontario, and University of Toronto, Toronto, Ontario, Canada; [20]Department of Gastroenterology, Sestopali Fund for Gastrointestinal Cancer Prevention, Tel Aviv, Israel; [21]Research Center for Cancer Prevention and Screening, National Cancer Center, Tokyo, Japan; [22]Department of Medicine and Epidemiology, University of Pittsburgh, Pittsburgh, Pennsylvania; [23]National Health Service Bowel Cancer Screening Southern Program Hub, Royal Surrey County Hospital, Guildford, United Kingdom; [24]Department of Surgery, University of Dundee, Dundee, Scotland; [25]Office of the Vice Chancellor, The Chinese University of Hong Kong, Shatin, China; [26]Gastroenterology and Nutrition Service, Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, New York

†Deceased.

We are grateful to the secretariat of the World Endoscopy Organization and to Alexandra Whibley, Erin Symonds, and Emma-May Palmer for their assistance in the consensus processes and article preparation.

This article is dedicated to the memory of Professor Paul Rozen, a true pioneer in colorectal cancer screening, who passed away in early 2013.

Members of the original working party responsible for the consensus process underlying these recommendations were convened by the World Endoscopy Organization Colorectal Cancer Screening Committee and were: G.P. Young, J. Mandel, J.J.Y. Sung, J.E. Allison, W. Atkin, R. Benamouzig, G. Hoff, S.H. Itzkowitz, T.R. Levin, E.G. McFarlane, C. O'Morain, S. Parry, L. Rabeneck, P. Rozen, H. Saito, R.E. Schoen, C. Senore, R.J.C. Steele, S.J. Winawer, and B.C.Y. Wong. The listed authors were members of the World Endoscopy Organization/World Gastroenterology Organization New Screening Tests Expert Working Party and met the criteria for authorship of this article.

Additional supporting information may be found in the online version of this article.

DOI: 10.1002/cncr.29865, **Received:** September 2, 2015; **Revised:** November 14, 2015; **Accepted:** November 30, 2015, **Published online** February 1, 2016 in Wiley Online Library (wileyonlinelibrary.com)

INTRODUCTION

New tests to screen for colorectal cancer (CRC) continue to emerge and are based on new biomarkers, new imaging modalities, or variations of existing methods. Efficient evaluation of these options presents a challenge. It has been observed that new *diagnostic* tests frequently enter practice without evidence of improved outcomes.[1] For *screening* tests, the requirement for evidence is more demanding, because more than clinical test accuracy (ie, sensitivity and specificity) is required to justify adoption.[1,2] Safety, public acceptability, and cost effectiveness need to be assessed even more carefully for tests that are to be applied to ostensibly healthy individuals.

The intention of a cancer screening program, or secondary prevention, is to significantly reduce the cancer site-specific mortality and burden of that disease in the target population[2] through programmatic use of a test that detects neoplasia at a stage early enough for treatment to be successful and/or for incidence to be reduced.[3]

It has been demonstrated that certain screening tests reduce cancer site-specific mortality and/or incidence by randomized, population-based evaluation on an intention-to-screen basis,[4-12] thereby limiting biases, such as lead-time, length, and self-selection, that are often present in simpler studies that use surrogate measures of mortality or intermediate endpoints. Evaluation of every new CRC screening test to the endpoint of mortality would be a huge and expensive undertaking and would markedly slow—if not prohibit—the implementation of promising new technologies. Fortunately, simpler studies using surrogate measures or intermediate endpoints can be used to evaluate new tests[1] provided that a carefully validated reference standard is used and biases are minimized. To define what is justifiably required to support the use of a new test for CRC screening, we propose an efficient and rigorous method for how to compare the alternative/new (hereafter "new") with the proven/established screening tests.

METHODS

To establish the guiding principles for comparative evaluation, including the informative endpoints and the appropriate study design, we established a consensus based on the Glaser and Delphi approaches.[13] The *membership* was chosen from experts because of their knowledge or experience in practice or research relevant to screening for CRC. The *problem* was defined by using the consensus process to agree on the goal. To support the consensus process, systematic literature searches were undertaken using Medline and other relevant databases. One search string was optimized for diagnosis and screening with the inclusion of measures like sensitivity, another was optimized for cancer, and a third attempted to identify articles focused

BOX 1: These are the guiding principles that underpin a strategy for comparing screening tests that emerged from the consensus approach and the literature review (a discussion of each is provided in Supporting Table 1; see online supporting information):

Principle 1. Screening aims to reduce the burden of disease in the targeted population, without adversely affecting the health status of those who participate in screening, through early detection and treatment of cancer and/or through detection of precancer lesions, which reduces incidence.

Principle 2. The screening test is just 1 event in a process that includes engagement of the public, testing, validation, communication, and treatment.

Principle 3. Population randomized controlled trials with mortality as the primary outcome set the standard for the evaluation of new tests.

Principle 4. New tests can be assessed in parallel with an existing test all the way through the screening process, from population engagement to population outcomes/measures.

Principle 5. New screening tests might detect a different neoplasia-dependent biology; as a consequence, the value of treatment and benefit to mortality reduction might not be the same.

Principle 6. In 2-step screening, the screening test selects participants who proceed to diagnostic verification by colonoscopy, because a positive test increases the likelihood of neoplasia being present.

Principle 7. It is not ethically justifiable to proceed to study a test in the screening environment, including acceptability to invitees or other screening program outcomes, without studies indicating that the new test is of acceptable accuracy compared with a proven comparator test.

Principle 8. New tests must be clearly defined with provision of adequate technical details, quality-assurance procedures, and performance standards.

TABLE 1. Characteristics of Established Screening Tests Known to Reduce Colorectal Cancer Mortality and the Type of Evidence Supporting Their Value

Detection Goal	Technology	Strongest Evidence for Benefit	Test Objective	Sensitivity Determinant	Specificity Determinants
Fecal blood	Guaiac-based FOBT (gFOBT)	Population RCTs—reduced incidence and mortality	Heme component of hemoglobin	Amount of fecal heme exceeds that needed to generate a positive result (fixed by manufacturer)	Dietary peroxidases; agents interfering with peroxidase reaction; bleeding nonneoplastic lesions; amount of stool in sample.
	Fecal immunochemical test for hemoglobin (FIT)	Case-control and cohort studies—reduced incidence and mortality; comparative screening cohorts (randomized)—higher detection rates and participation compared with gFOBT	Globin component of human hemoglobin	Amount of fecal hemoglobin exceeding selected cutoff concentration (may be fixed by manufacturer or selected by end user)	Bleeding nonneoplastic colonic lesions; amount of stool in sample.
Endoscopic visualization of lesion	Colonoscopy	Case-control and cohort studies—reduced incidence and mortality	Visually apparent lesions (ulcerative, polypoid, or flat/depressed) suspicious of neoplasia	Quality of procedure; ability to negotiate the colonic lumen with adequate views; nature of the lesion	Histopathologic clarification
	Sigmoidoscopy (flexible)	Population RCTs – reduced incidence and mortality	Visually apparent lesions within reach	Quality of procedure; depth of insertion; ability to negotiate the colonic lumen with adequate views; nature of the lesion	Histopathologic clarification

Abbreviations: FOBT, fecal occult blood test; RCTs, randomized controlled trials.
[a] This information is derived from several publications.[5,6,14-19]

on comparison of tests. We also searched for review articles that addressed the evidence supporting screening for CRC.

A series of *specific questions* that focused on the definition of appropriate study designs and outcomes for the comparison of different screening tests were established by agreement. The *answers* to these questions were reached by consensus (requiring 75% agreement) based on dissemination of summaries of the literature searches, detailed examination of methodological articles, a series of semistructured discussions with dissemination of decisions after each critique, followed by consultation with external advisors. On the basis of these processes, progressive drafts of the recommendations were then prepared, circulated, and critiqued.

In this report, we present: 1) the underlying guiding principles that emerged from the consensus; 2) an expert opinion on the methods appropriate for evaluating a new test compared with a proven comparator test (what is needed), 3) practical guidance on how to apply these methods in a 4-step, phased evaluation (how to do it); and 4) examples of published research that exemplify these phases (how it has been done). Therefore, it will guide researchers and enable practitioners to decide whether a new test is suitable for the context in which they practice.

GUIDING PRINCIPLES

The guiding principles that emerged from the consensus approach and the literature review are outlined in Box 1, together with their key consequences for test comparison. A presentation of the reasoning underlying these principles is presented in Supporting Table 1 (see online supporting information).

With regard to Principle 3, which states that "Population randomized controlled trials (RCTs) set the standard for evaluation of new tests," Table 1 outlines the characteristics of major screening tests known to reduce CRC mortality and/or incidence together with the type of evidence supporting their value. Such tests are ideal as a reference point against which to compare a new test.

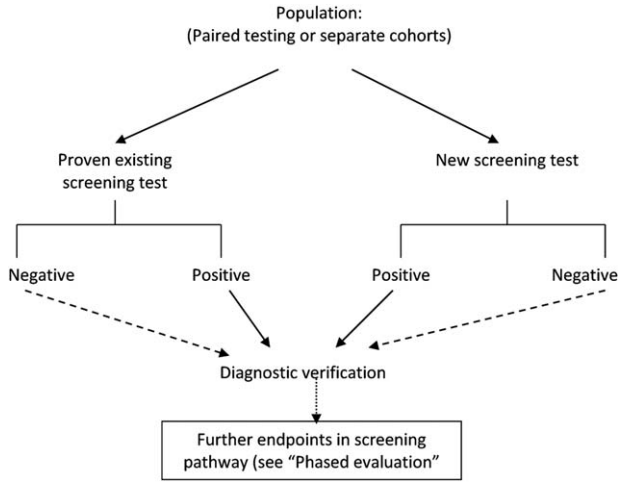

Figure 1. This is a conceptualization of the design for testing a new test relative to an existing (comparator) test. Solid lines represent essential paths in the process, and dashed lines represent discretionary paths that are not essential in some phases of evaluation.

Table 1 also describes the test target (which serves as an informative outcome for comparison), as discussed in Principle 5.

A FRAMEWORK FOR EVALUATING A NEW SCREENING TEST

With these principles in mind, a practical framework for evaluating a "new" test against a proven test can be built. The test of effectiveness for the proven test demands proof at the population level—hence, the context for evaluation must eventually include population outcomes and not just the testing of capacity to detect lesions.

When an RCT establishes that a test is effective in reducing mortality, then a new test does not need to be evaluated with such rigor provided it is compared with the proven test.[1] This is true provided that Principle 5 (Box 1) applies; namely, that the value of treatment and benefit in mortality are not compromised because of potential differences in the biology of detected lesions.

In applying this view, other than effects on CRC mortality and disease stage, there are 3 types of readily determined outcomes that inform the value of a new test: accuracy, acceptability, and impact on other screening program outcomes when applied in a screening context (see Phased Evaluation, below). Such intermediate/surrogate outcomes facilitate the prediction of benefit provided that the new test is directly compared with a test that has been proven to be effective on an intention-to-treat basis, ie, based on an approach that, among other things, takes into account imperfect adherence and overcomes other sources of bias.[1,20,21]

STUDY DESIGN FOR COMPARING TESTS

Accuracy can be assessed through case-control and cohort studies using the framework illustrated in Figure 1. This framework can be adapted to any phase of evaluation, from prescreening assessment to mass population application.

Choice of Comparator Test

The first and well characterized, noninvasive test (in terms of effectiveness) is the guaiac-based fecal occult blood test (gFOBT) Hemoccult (and variants, particularly Hemoccult II; Beckman Coulter Inc, Pasadena, Calif). The screening outcomes achieved with this gFOBT represent the *minimum* that needs to be achieved, because the effect of gFOBT on mortality is modest. The more advanced technology provided by fecal immunochemical tests for hemoglobin (FIT) provides better accuracy, including improved sensitivity for adenomas as well as CRCs and better acceptability when evaluated on an intention-to-screen basis. Population-based and case-control studies support the value of this technology.[22-29] Further studies from the Netherlands[16] confirm the value of FIT in a population RCT when analyzed on an intention-to-screen basis relative to the gFOBT Hemoccult II. This evidence has led to recommendations that FIT replace gFOBT.[15,30] Therefore, a *well studied* FIT sets a new standard against which new tests can be judged.[31] FIT technology tends to have a better capacity to detect adenomas than gFOBT, and repeated testing improves detection.[32,33]

Because population screening trials with flexible sigmoidoscopy (FS) have now been reported,[5] this screening test will serve as a useful comparator for the detection of preinvasive lesions.

The experts concluded that colonoscopy serves to estimate the accuracy of a new test; however, without RCT intention-to-screen evidence of effectiveness, the effectiveness of a new noninvasive test cannot be deduced if it is assessed relative to colonoscopy only. However, as results emerge from the currently underway population screening trials evaluating colonoscopy, we will be able to use colonoscopy as a comparator knowing its benefit to mortality in an unbiased setting.

EVALUATION OF ACCURACY

Clinical accuracy (sensitivity, specificity, and predictive values) is crucial to whether a new test is fully evaluated in

TABLE 2. Relation Between Direct Practical Measures (Operating Characteristics) of a Screening Test Result, How Each Informs Assessment of Test Accuracy, and the Consequences of the Result for a Screening Program

Test Result	Diagnostic Verification; Operating Characteristic	Corresponding Accuracy Characteristic	Issue Addressed
Positive	True (ie, target condition present); true-positive rate (TPR)[a]	Sensitivity (positivity rate in those with the target condition)	Detection
		Positive predictive value (TPR/TPR + FPR)	Efficiency of detection
	False (ie, target condition not present); false-positive rate (FPR)[a]	Specificity (1 − FPR)	Burden associated with detection
Negative	True; true-negative rate (TNR)	Negative predictive value (TNR/TNR + FNR)	Elimination/exclusion of targeted clinical lesion (stage specified)
	False; false-negative rate (FNR)	Missed lesion	Burden of failed detection

[a] A targeted clinical lesion is either cancer and/or advanced adenoma, depending on the question being asked of the test, because tests might detect these to differing degrees.

screening.[1] It is not appropriate to study acceptability or other screening program outcomes without having first measured accuracy. Consequently, comprehensive test evaluation must be phased (see Principle 7).

The 2 key measures of accuracy—sensitivity and specificity—are often difficult to ascertain, especially for screen-relevant lesions (ie, the earlier stage cancers and adenomas that would be encountered in a largely asymptomatic, typical screening population). A valid estimate of these accuracy measures would require costly and time-consuming testing of an unselected screening population that included a sufficient number of participants with such lesions in which all test confounders were likely to be encountered and in which every participant, both test-positive and test-negative, underwent diagnostic verification.

Fortunately, when a comparator test is available, a paired study design (which improves statistical power) facilitates evaluation of effectiveness of the new test and estimation of the *relative* impact on screening outcomes. We conclude, in line with others,[21] that existing tests, namely gFOBT/FIT and FS, have demonstrated effectiveness and can be used to facilitate assessment of relative benefit.

Another simplification is based on the proposition that the 2 key questions concerning clinical accuracy[3,34,35] are: 1) detection—a test that is more sensitive in practical terms returns more true-positives, and 2) the burden associated with detection— a test that is more specific in practical terms returns fewer false-positives. *The assessment of these 2 parameters is achieved by a thorough diagnostic verification of every test-positive case (both comparator and new test-positives) to determine whether it is a true-positive or a false-positive.*[3,36]

The simple dichotomous measures of the true-positive rate (TPR) and the false-positive rate (FPR) are direct and practical measures of accuracy, sometimes referred to as test "operating characteristics," as indicated in Table 2. They are used when undertaking receiver operating characteristic (ROC) analysis. The TPR reflects detection (sensitivity), and the FPR reflects the burden associated with detection (1-specificity). Consequently, relative sensitivity and specificity are determined by comparing the TPR and the FPR, respectively, between tests.

Comparing Test Accuracy: The Scenarios

The approach based on verification of positive tests, classifying them as true-positive or false-positive, provides a straightforward but powerful strategy for comparing the accuracy (operating characteristics) of 2 screening tests. The concepts presented apply regardless of whether the target lesion is cancer and/or adenoma.

In comparing accuracy, the targeted clinical lesion (hereafter referred to as *targeted lesion*), which can be cancer, and/or adenoma, or combinations thereof, needs to be clearly defined. Performance characteristics related to sensitivity and specificity need to be compared for the same clinical endpoint. Depending on the phase of evaluation and the question being addressed, the target lesion might be early stage cancer, or advanced adenoma, or "advanced neoplasia," a term referring to cancer plus advanced adenoma (see Phase 2 below for definition). Tests might differ in their capacity to detect lesions at specific stages, and this needs to be explored. It should be noted that clinical accuracy depends on the presence of the biomarker that forms the basis of the test objective (see Table 1); and this, in turn, might be important to

treatment response (Principle 5) (Supporting Table 1; see online supporting information).

Two simple questions, modified from Lord et al,[1] guide assessment in a practical manner:

Is the new test better at detecting target lesions?

This is true if the TPR (which reflects sensitivity) for the target lesion is improved using the new test. It is likely that improved outcomes (reduced mortality and/or incidence) will follow from use of the new test, especially if the TPR is greater for early stage cancers.

Complexity arises if the new test is better at detection (higher sensitivity) but returns more false-positives (lower specificity) than the old test, raising concerns about cost and potential harms. Hemoccult Sensa, compared with Hemoccult II, is an example.[37,38] Note, however, that a test with more true-positives and a higher initial colonoscopy rate (whether because of true-positives and/or false-positives) will make the program more expensive initially but might create longer term savings as a result of better detection. This will become clearer in formal cost-effectiveness analyses that measure the cost per quality-adjusted life year saved.

There are several ways to address such complex scenarios. The operating characteristics of the 2 tests can be plotted as an ROC curve (TPR vs FPR) as a way to judge which test has the best balance of true-positives and false-positives; overall, the test with the greatest area under the curve has the best discriminatory power.[39] This is particularly applicable to prescreening phases in the evaluation process that focusses on accuracy (see below).

Another objective approach is to calculate the number needed to screen (colonoscope) to detect 1 target lesion using each test (the reciprocal of the positive predictive value). Calculating the number needed to colonoscope also facilitates comparison of 2 tests when each is applied to a different cohort, although comparability of populations needs careful consideration. However, the number needed to colonoscope should be determined only in Phase 3 studies conducted in settings that represent the natural prevalence of neoplasia and not in studies in which prevalence is biased because of recruitment processes.

If not better at detecting target lesions, does it have other advantages?

A new test might have other benefits, for instance, significantly better specificity without improved sensitivity. Comparison is made simple in this circumstance by calculating the number needed to colonoscope to detect 1 target lesion for each test. The new test might also have programmatic benefits (see Phase 3 evaluation), such as greater acceptance by the screening population or improved technical reliability. In similar fashion, the number needed to invite to detect 1 target lesion will offer additional comparative information by capturing the product of participation and accuracy, although this approach is susceptible to the method of invitation and how the invitation is framed. It should be noted that many consider the sensitivity of gFOBT, which has demonstrated a statistically significant but only relatively small impact on CRC mortality, to be inadequate. Consequently, they would argue that there is only a place for a new test that returns a better sensitivity than gFOBT.

Study Populations

The population selected for study will depend on the question being asked and the phase of the evaluation. The testing *path* may involve paired testing in a single group (that comprises cases and controls) or parallel testing of randomized cohorts (see Fig. 1). Which is chosen depends on the stage of evaluation (see Box 2). The subsequent discussion on phased evaluation provides more detail.

COMPARING TESTS IN THE SCREENING PATHWAY

In addition to accuracy, it is essential that the effect of a new test on other variables in the screening pathway is determined, eg, safety, cost, feasibility, ease of use for a screening participant, and acceptability. New tests must undergo evaluation in unselected, typical screening populations, and an intention-to-screen evaluation is necessary to justify large-scale adoption.

In mass population screening, detection of target lesions is the product of participation and sensitivity; because, without participation (sometimes referred to as compliance or uptake), there can be no detection.[41] Consequently, measuring participation with 1 test relative to another in separate cohorts randomly selected from the same population can document test acceptability,[42] provided that framing of information is carefully balanced.

PHASED EVALUATION

Phased (ie, sequential) evaluation in a step-wise, increasingly complex manner is most appropriate.[3,20,43-46] Initial evaluation (Phases 1 and 2) starts with a simple prescreening evaluation that addresses accuracy of the new test and proceeds, if judged appropriate, to more

BOX 2: Study Populations and Testing Path

- **Initial testing of accuracy (Phases 1 and 2):** Ideally a single clinical group of patients undertaking *paired* testing (ie, each does both the new test and the old test), as shown in Figure 1. This is an efficient design. Initially, diagnostic verification of all cases by colonoscopy is carried out regardless of test results. Pairing reduces cohort size because of improved statistical power for assessing incremental benefit. It ensures that individuals are comparable and avoids imbalances in variables that affect test results and in other biases between the tests. If the new test demonstrates promise, then larger numbers of individuals undertaking paired testing can be further studied with colonoscopic follow-up in test-positive individuals only.
- **Subsequent testing in the screening context (Phases 3 and 4):** Individuals may be randomly assigned to do either the proven or the new test, in the context of the screening pathway, on an intention-to-screen basis, when it has been demonstrated first that the accuracy of the new test is not worse than that of a suitable, proven comparator test. When assessing test accuracy in parallel groups, the inclusion criteria for the study group must be carefully characterized and the detected lesions fully described. Without this, transferability from 1 setting to another is not possible.[40]

BOX 3: The 4 Phases of Test Evaluation and Associated Issues

Phase 1. Retrospective estimation of ability to discriminate between cancer cases and normal;
Phase 2. Detection of presymptomatic stages along the neoplastic continuum, prospective clinical studies;
Phase 3. Initial screening evaluation—participation and prevalence studies; and
Phase 4. Screening program evaluation.
Issues to be noted:

- In 2-step screening, screening tests select participants[21] who then undergo the reference diagnostic test.
- Pathway parameters in screening, such as participation rates, are as crucial to population benefit as accuracy.[41]
- Relative test accuracy is simply addressed in a paired design.[3]
- The value of the new test should be compared with the old test in the context of how the new test is to be implemented in the existing screening pathway.[21]
- The specific phases of screening are a guide to evaluation reflecting a continuum from simple to increasingly complex evaluations in which each step may be adjusted for complexity according to outcomes in the previous phases.[21]
- The cost for each phase is subject to local considerations; however, if the costs of diagnostic verification are put aside, then Phase 1 studies might cost several hundred thousand dollars, whereas Phases 3 and 4 will cost several to many millions of dollars.

thorough evaluation addressing outcomes in the population screening context (Phases 3 and 4), as indicated in Box 3. Phased evaluation takes into account the issues described in Box 3. The primary and secondary objectives and general characteristics of these phases are provided in Table 3.

The phased approach is indeed undertaken in practice. Supporting Table 2 (see online supporting information) provides selected examples of studies that have demonstrated the elements of each phase, together with their main characteristics. When tracking through these phases for tests, such as different FIT products and designs or the fecal DNA tests, it is observed that early, simple studies are followed by more complex and informative studies. There are many other possible examples—those provided serve to demonstrate the increasing complexity of each phase, design options within a phase, and information that may be gleaned from such studies.

Phase 1: Retrospective Estimation of Ability to Discriminate Between Cancer Cases and Normal Controls

The ability to distinguish between cancer and noncancer states is essential for a test to be useful and can initially be evaluated in individuals who have established cancer

TABLE 3. Phased Evaluation for Comparison of Screening Tests for Colorectal Cancer[a]

Evaluation	Nature	Primary Aim	Secondary Aims	Population
Phase 1	Prescreening: Retrospective estimation of ability to discriminate between cancer cases and controls without neoplasia	• Test detects established cancer 1.1 To estimate TPR and FPR (test operating characteristics) as the primary measures of accuracy relative to an established test	1.2 Establish the test sampling process 1.3 Optimize processes for quality assurance 1.4 Fine tune test endpoint	Individuals known to have cancer, ideally with a majority in potentially curable disease stages and including some who are asymptomatic; controls to be free of neoplasia; concordance between tests should be reported; ideally, paired testing, with all results verified at diagnostic procedure
Phase 2	Detection of lesions along the neoplastic continuum; prospective clinical studies	• Test detects early neoplasia before it becomes apparent 2.1 To estimate test operating characteristics for detection of neoplasia at stages along the oncogenesis continuum, especially preclinical disease, including advanced adenomas 2.2 To determine the final format of the test (sample and endpoint) Minimum requirement for test registration	2.3 More reliably estimate operating characteristics 2.4 Information on covariates affecting test performance 2.5 Ascertain the number of samples and threshold (fine-tune the endpoint) 2.6 Test to be registerable with authorities 2.7 Clarify whether there are subgroups in which the test might fail to detect lesions	Cases covering all stages of colorectal neoplasia, especially early stage cancer and/or advanced adenomas, with knowledge of whether cases are symptomatic; asymptomatic where possible; controls to be free of neoplasia; results in individuals with common benign diseases and how they affect test result need ascertainment; testing undertaken before scheduled diagnostic procedure; ideally, paired testing; concordance between tests should be reported
Phase 3	Initial screening evaluation; single round of screening	• Characteristics of neoplasia detected when screening; false-referral rate; acceptability 3.1 In a screening population, to determine the operating characteristics of the test, what is detected, and the workload associated with detection, including the false-referral rate. 3.2 Determine test acceptability Minimum requirement for use in organized screening	3.3 Describe the characteristics and frequency of neoplasia detected when screening 3.4 Determine feasibility 3.5 Preliminary assessment of costs including diagnostic workload	Testing in a typical screening environment using a single prevalent screen; separate cohorts perform the new test or comparator (potentially in the form of "usual care"), and outcomes are followed from invitation to outcome of interest; only those who test positive need colonoscopy (unless direct comparison with screening colonoscopy is required); start with initial, small studies addressing simpler pathway outcomes and progress to larger programs addressing detection rates; analyze by intention-to-screen
Phase 4	Screening program evaluation over multiple rounds	• Impact of screening on reducing burden of neoplasia, adverse events 4.1 To estimate or model reductions in cancer mortality	4.2 Broader benefits 4.3 Accurate costs 4.4 Participation with rescreening 4.5 Compliance with diagnostic follow-up 4.6 Treatability of lesions detected 4.7 Screening intervals 4.8 Missed cancer rate 4.9 Program detection rates with repeated screening 4.10 Diagnostic follow-up rate across all rounds 4.11 Number needed to screen to detect a lesion 4.12 Unexpected adverse events	Randomly selected from populations in which screening program is likely to be implemented; design may use historic controls or else a parallel-arm RCT with screening participants and alternatively screened population; intention-to-screen analysis required

Abbreviations: FPR, false-positive rate; RCT, randomized controlled trial; TPR, true-positive rate.
[a] Discussions of group sizes and approximate costs for each phase are included in the text.

(cases) compared with those who are free of neoplasia (controls). Although they initially guide evaluation, the accuracy measures obtained in this way may be biased, and the cases used are not necessarily representative of preclinical cancer, the critical target of any screening program.

Cases and controls

An initial indication can be obtained comparing individuals who have established cancer (cases) with those who are free of neoplasia (controls). For cases, it is helpful to have a range of different histologic features and stages, meaning that all must have had diagnostic colonoscopy.

Intervention

Design should follow that charted in Figure 1, with cases and controls performing both the new tests and the comparator tests: ie, "paired-testing." The individuals who are developing the test sample should be blinded with respect to participants' status. If the test requires collection of biologic samples, then it needs to be ensured that the sampling process and preanalytic conditions are exactly the same for cases and controls (such as time interval from the colonoscopy, setting of the examination, conditions of sample storage, and so on).

Outcomes and sample size

A sample size of 60 pairs has approximately 80% power to detect a difference in the TPR of 20% when the proportion of discordant pairs is expected to be 30% in cases affected by the cancer; such conditions may be encountered.[47] "Discordant pairs" refers to those cases who are positive on 1 or the other test but not on both tests. The minimum standard approach and its analysis is described in detail by Pepe et al.[3] Basic considerations in measuring power when the TPR and the FPR are the main outcomes and when the design is not paired have been provided.[3]

For studies on marker combinations that require training before validation, if the training and validation cases are drawn from the same population, then the sample size requirements should be fulfilled by the validation set independent of the training set.

The proportions of individuals with lesions in which both the new test and the comparator test are positive and in which only 1 or the other test is positive should be reported. This clarifies concordance between the tests and addresses Principle 5.

To compare tests in a paired design, calculation is simply performed by determining the confidence interval of the difference in test positivity[8] or by using the McNemar test. Fine-tuning the test endpoint, ie, the threshold set for positivity (the *criterion* value), is crucial for those tests that have a quantitative or semiquantitative endpoint. An ROC curve should be constructed and analyzed.[3,39] For each cutoff selected for positivity in the ROC curve, the confidence interval of the difference in positivity rates between the new test and the comparator test can be calculated.[47]

If the new test is at least comparable to the comparator test, then it is justified to proceed to a Phase 2 evaluation. In exceptional circumstances, skipping phases before Phase 3 might be justifiable, especially if screen-detected cases were included.

Phase 2: Detection of Neoplasia Across the Oncogenic Continuum—Prospective Clinical Studies

Paired testing is undertaken prospectively in participants before they undergo the diagnostic procedure: ie, before they are identified as cases or controls. Test operating characteristics need to be understood across the spectrum of stages of oncogenesis, with the particular interest being performance in the earlier stages, when treatment is more likely to be successful. This is especially important if the new test has a different objective (ie, it detects a different biology) than that of the proven comparator. The risk in practice is that seeking a higher detection rate for early stages or preinvasive neoplasia (adenomas) raises the possibility of a higher FPR and overdiagnosis (detection of inconsequential colorectal neoplasia).[48]

There are 2 clinical targets of particular interest. One is a shift to earlier stage cancer, because CRC screening RCTs demonstrate that reduced mortality is linked to earlier detection. This can only be examined in very large screening studies,[49] but a surrogate measure is provided by estimating sensitivity for earlier stage cancer. The second target is that of preinvasive neoplasia, particularly *advanced* adenomas (size >9 mm, villous component >25%, high-grade dysplasia, or >2 of any characteristic), because the detection of adenomas by screening FS is beneficial,[5,50,51] and advanced adenomas are more likely to progress to cancer.

An important purpose of Phase 2 can be to determine the final test format (ie, criterion endpoint fine-tuning), before the population evaluation in Phase 3. The operational nature of the test (eg, in the case of a laboratory test, the assay details and analyte) should be carefully defined (see Principle 8), and a provisional threshold should be set for positivity: ie, the characteristic that would direct that individual to undergo diagnostic evaluation. For tests requiring a biologic sample, the sampling process must be clear; information on stability of the

analyte and robustness of the sampling method regarding preanalytic variations should be published. If any of these matters remain uncertain, then simple pilot studies in typical screening populations should be undertaken. Although a new test might detect lesions at an earlier stage, it also might fail at certain stages, or it might detect a different type of neoplastic lesion. Ideally, Phase 2 studies would indicate whether these outcomes are likely.

Cases and controls

Individuals who are scheduled for colonoscopy for any reason are informative, but they are more so if asymptomatic.

Intervention

Evaluation parallels that for Phase 1, with individuals undergoing paired testing before colonoscopy. Participants should be classified according to stage of oncogenesis and presence or absence of neoplasia, specifically: cancer stage, advanced adenoma, nonadvanced adenoma, benign pathology, or normal organ.

Generalized linear modeling can be used to examine the relation between covariates and test results.[36,42] This will highlight the factors other than pathology in the organ that must be considered in Phase 3 as potential covariates.

Outcomes and sample size

The low prevalence of cancer, even in individuals who are scheduled for colonoscopy, requires the recruitment of many participants. A meaningful comparison may be achieved if approximately 60 of the desired target lesions are included in the study population given paired-testing, as discussed for Phase 1. To calculate the total population size required to provide sufficient power, the likely prevalence of the target lesion in the population must be known. From 1000 to 5000 individuals should be recruited as a general rule, depending on whether attempts to enrich the population with cancer cases are successful. Advanced adenomas are likely to be ascertained at a rate approximately 3 to 10 times that of cancer when evaluating screening tests for CRC.

The data provided from Phase 2 evaluation may be sufficient to have a test registered with appropriate authorities for medical use. If performance has been demonstrated to be at least equivalent to that of the comparator, then it is justifiable to proceed to population screening studies.

Phase 3: Initial Screening Evaluation—Participation and Prevalence Studies

Phase 3 evaluation seeks to confirm that the new test improves outcomes when the test is applied in the screening context as a 1-time event: ie, a prevalent screen. Usually, separate cohorts are randomized to each test to provide intention-to-screen outcomes. An organized screening program starts with an offer of the test, the test sample is obtained by the participant (ideally under optimal conditions) but entirely at their own discretion, the sample is submitted for analysis, and each positive test result must be verified by a diagnostic test.[52] *This is the minimum level of evidence required to justify use in large-scale, organized screening.*

The population

Study groups should be derived randomly from a population that would be targeted in a screening program. Unbiased selection of invitees is highly desirable.

Intervention

In randomized screening trials, participants usually perform 1 test only, as though this were a typical screening program. If they do both, then intention-to-screen outcomes cannot be determined. Prospective testing with either the new test or the comparator test requires that sample collection is undertaken before ascertainment of the diagnosis. Events should be tracked from the offer of screening to the completion of diagnostic verification (see Principle 4), except in small studies that seek to gather information on participation as the only outcome.

Outcomes and sample size

Both an intention-to-screen analysis of results and a per-protocol analysis should be undertaken. For per-protocol (ie, participant) analyses, in addition to the outcomes discussed above, the overall test positivity rate, which defines the total diagnostic workload (ie, colonoscopy), is informative. For intention-to-screen analyses, test participation rates and tracking the return of tests over time are also informative.

Adjusted logistic regression analyses can be undertaken to adjust for covariates.[36,42] Because separate groups are studied in this type of design, covariates may not be equal between the groups, and they especially might not be equal between those undertaking testing or returning positive test results.

Sample size depends on the degree of incremental improvement being sought, the target lesion of interest, whether the focus is on an intention-to-screen or

participatory (per-protocol) outcome, and the outcome being addressed. For instance, test positivity or participation rates are often the initial outcomes of interest in Phase 3 studies and are easily estimated. With study group sizes of n = 376, a 2-group chi-square test with a .05 two-sided significance level has 80% power to detect a 10% change in participation, where participation in the reference group is 30%.[42] When the ultimate consideration is the difference in detection rates of cancer, if a difference in detection rates of cancer of 3 per 1000 invitees is expected,[7] then the sample size should be at least 6083 if a gFOBT comparator is expected to detect 2 per 1000.

Therefore, it is sensible within Phase 3 studies to progressively stage evaluation, starting with smaller study groups of, say, 400 to 500 to measure the overall test positivity rate (which estimates the number of colonoscopies required to be) and participation rates and to gain further estimates of the TPR and FPR and associated covariates. This informs sample sizes for larger studies that then address detection rates. Modeling cost effectiveness is an important element of Phase 3, because it provides real-world estimates of test positivity rates and participation, variables that are important to accurate cost modeling. Indeed, as outcomes are accumulated, extensive modeling can be undertaken using models like MISCAN (Microsimulation Screening Analysis)[53] to predict impact and thus enable the adjustment of programs to maximize the likely benefit.

Phase 4: Screening Program Evaluation

The objective of screening is to reduce the burden of disease by reducing CRC mortality at the population level. It is important that it does not adversely affect the health status of those who choose to participate. A new test might be associated with some unexpected adverse events that would counterbalance mortality benefits predicted by better detection and/or participation; Phase 4 studies conducted over multiple rounds should identify these events.

Comparing new CRC screening tests using CRC mortality as the endpoint will probably never be feasible on the grounds of size, time, and cost. Phase 4 evaluation is not so much about the comparison of tests but about monitoring how the new test performs when applied to a large, unselected population, ideally over repeated rounds of screening. Measures like a shift to an earlier disease stage and interval (missed) cancers are ascertainable, as well as unexpected adverse events. Knowledge of these will improve cost-effectiveness determinations. Consequently, Phase 4 evaluation would normally proceed as a process of careful evaluation of an organized screening program applied to a large population and monitored over a considerable time, often involving multiple rounds of screening.

Outcome measures that demonstrate benefit

In considering what to measure to assess health benefits in screening programs, intermediate measures associated with demonstrated RCT effectiveness can be informative.[14] The gFOBT RCTs demonstrate that a shift to an earlier stage of cancer in a program that involves repeated screening offers is associated with reduced mortality.[7,8,10,54] Thus earlier detection by a new test to at least a comparable degree is highly desirable; for instance, it has now been demonstrated that screening with FIT leads to earlier detection.[49]

The association of adenoma detection and removal in screening with the reduction of CRC incidence and mortality is now proven by the RCTs of FS screening.[5] Thus FS is an expeditious comparator for evaluating new tests that target preinvasive lesions, because a potential surrogate measure for predicting a reduction in incidence is the detection (the TPR) of those lesions considered to be at high risk of progressing to CRC.

Interval cancers, ie, missed or new cancers, occur in programs, and monitoring these for each test would be valuable; although, to obtain valid and accurate comparative data, an adequate follow-up time and a very large sample size are required. Nonetheless, interval cancer rates need to be determined, especially when the earlier phases of evaluation have focused primarily on assessment of test-positive cases (ie, an endoscopic method is not routinely undertaken in test-negative participants).

Comparing tests over multiple rounds is also an important goal of Phase 4 testing and will require prolonged follow-up. Cumulative detection rates should be considered when the stipulated screening interval of the tests being compared is different. Also, methods for reporting participation over multiple rounds of screening have not been well applied to CRC screening[55]; however, as long as repeated participation is required to achieve the expected screening benefit, this represents a relevant indicator to be assessed. Participation in screening—a central performance indicator for population screening—can vary across the population, and it is important to monitor not only the effect of a new test on overall uptake but also its acceptability to all socioeconomic and ethnic groups to avoid widening the inequalities gap.

Phase 4 study design

Studies should follow the design outlined for Phase 3 evaluation but should also include multiple rounds of

screening (at least several with the interval matched to the perceived duration of effect of each test), with plans to ascertain the outcomes relating to those measures deemed important; namely, participation, detection, cost, adverse effects, earlier detection, and interval (new or missed) lesions.

Such studies will be extremely costly and normally would be feasible only in the context of public health screening strategies that are already in place, in which methods to collect outcome measures are already designed and operational. In other words, Phase 3 evaluation is sufficient to lead to the incorporation of a new test into a pilot within a formal, organized population program, and Phase 4 evaluation serves to confirm the expected promise by an evaluation of screening programs. Given good information on costs, the comparative cost effectiveness of different tests can be determined as described.[56]

NEW BIOMARKERS

The discovery of new biomarkers, such as fecal or blood tests for DNA, RNA, or protein, adds complexity. Initial research usually precedes Phase 1[3] as we describe it but also requires fine-tuning the test endpoints in Phases 1 and 2. This is especially true if a panel of markers is being used.

The process of discovery starting with tissue banks has been discussed in detail elsewhere.[3,57] Sophisticated, retrospective molecular analyses of material in biospecimen banks can serve to identify candidate biomarkers that might become the objective of the screening test.

If such laboratory research identifies a promising biomarker, then it can be initially evaluated as for Phases 1 and 2 by a simple study in cases and controls. Doing this, however, may assume that the retrospective biospecimen banks are adequate to identify the best candidate. Usually, this is not the case, because discovery is often undertaken on limited numbers of samples obtained from strictly categorized materials that often are not typical of screen-detected lesions. A further technological challenge arises if resected tissue specimens are used to identify the biomarker; however, use of the biomarker in screening involves measurement in a biologic sample, such as blood or feces. Many factors may influence the appearance of the biomarker in the biologic sample, and there is a chance that it might not be of the same molecular structure in blood or feces as in tissue, because degradation or other processing might occur.

This makes it likely that the best discovery process first develops a putative panel of markers and then uses clinical studies set up in such a way that the panel can be explored in clinical specimens as part of Phase 1 or 2 studies, or perhaps even Phase 3 studies. Indeed, access to the appropriately characterized population with biologic samples, which serve as a source of materials for discovery of potential biomarkers, may be very useful. The usefulness of panels of multiple markers can then be explored, ie, "validated," in Phase 1, 2, and 3 studies.[57]

DISCUSSION

This phased approach provides an efficient method for evaluating a new screening test that increases in cost and complexity only if key attributes are worthwhile. It assesses both accuracy and acceptance, because screening of a general population requires good participation as well as good detection, and the same principles can be applied to adenoma detection.

Study costs increase considerably with each phase. The high cost of undertaking Phase 3 studies might be reduced by obtaining government regulatory approval for the use of a test on the basis of Phase 2 studies. Some authors suggest that this can wait until Phase 4 studies have been undertaken,[3] although that seems impractical, because no commercial entity would proceed with test development under such circumstances. Using the logistics and infrastructure of existing screening programs can also help reduce costs of such studies. Expensive studies have included the evaluation of new, noninvasive tests in colonoscopic screening participants.[58] Although useful, this fails to provide comparison with a test known to reduce mortality on an intention-to-screen basis.

The final issue is what justifies progression from 1 phase to the next. Although our proposal sets the principles for the phased evaluation of new tests, researchers, in collaboration with health service providers, should agree on hurdle values before embarking on a study. It is noteworthy that criteria for equivalence or superiority should be agreed at commencement. Phase 1 studies can be considered as exploratory and of value in helping to determine necessary power and likely outcomes in Phases 2 and 3. What constitutes an acceptable hurdle value will vary with the test and how the test will be used within the health care system.

We consider that this process of comparative, phased evaluation provides a rational, efficient, and useful process for evaluating new tests and for progressing a test to a stage at which the considerable degree of evidence needed for its inclusion in population screening is obtained. Health providers will be able to adopt a test that is soundly based on scientific objectivity and the fundamental principles of screening.

FUNDING SUPPORT

This work was supported by a grant from the World Gastroenterology Organization.

CONFLICT OF INTEREST DISCLOSURES

Graeme P. Young reports grants and nonfinancial support from Eiken Chemical Company and grants and personal fees from Clinical Genomics Pty Ltd outside the submitted work; he has a patent "Down Markers: A Method of Diagnosing Neoplasms (II PCT/AU2008/001565) licensed to Clinical Genomics. Carlo Senore reports nonfinancial support from Covidien/Given Imaging, Medical System, and Given Imaging outside the submitted work. Wendy S. Atkin reports nonfinancial support from Eiken Co. Ltd (MAST is UK distributor) outside the submitted work. Stephen H. Itzkowitz reports personal fees from Exact Sciences Corporation outside the submitted work. Bernard Levin reports personal fees from Exact Sciences ad Medial Research during the conduct of the study. Gerrit A. Meijer reports nonfinancial support from Exact Sciences and Sysmex outside the submitted work; he has a patent pending on biomarkers for the early detection of colorectal cancer.

REFERENCES

1. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Ann Intern Med.* 2006;144:850-855.
2. Wilson JMG, Junger G, World Health Organization (WHO). Principles and Practice of Screening for Disease. Public Health Papers, no. 34. Geneva, Switzerland, WHO; 1968.
3. Pepe MS, Etzioni R, Feng Z, et al. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst.* 2001;93: 1054-1061.
4. Atkin WS, Edwards R, Kralj-Hans I, et al. Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. *Lancet.* 2010;375:1624-1633.
5. Elmunzer BJ, Hayward RA, Schoenfeld PS, Saini SD, Deshpande A, Waljee AK. Effect of flexible sigmoidoscopy-based screening on incidence and mortality of colorectal cancer: a systematic review and meta-analysis of randomized controlled trials [serial online]. *PLoS Med.* 2012;9:e1001352.
6. Holme O, Loberg M, Kalager M, et al. Effect of flexible sigmoidoscopy screening on colorectal cancer incidence and mortality: a randomized clinical trial. *JAMA.* 2014;312:606-615.
7. Hardcastle JD, Chamberlain JO, Robinson MH, et al. Randomised controlled trial of faecal-occult-blood screening for colorectal cancer. *Lancet.* 1996;348:1472-1477.
8. Mandel JS, Bond JH, Church TR, et al. Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota Colon Cancer Control Study. *N Engl J Med.* 1993;328:1365-1371.
9. Mandel JS, Church TR, Bond JH, et al. The effect of fecal occult-blood screening on the incidence of colorectal cancer. *N Engl J Med.* 2000;343:1603-1607.
10. Kronborg O, Fenger C, Olsen J, Jorgensen OD, Sondergaard O. Randomised study of screening for colorectal cancer with faecal-occult-blood test. *Lancet.* 1996;348:1467-1471.
11. Segnan N, Armaroli P, Bonelli L, et al. Once-only sigmoidoscopy in colorectal cancer screening: follow-up findings of the Italian Randomized Controlled Trial—SCORE. *J Natl Cancer Inst.* 2011; 103:1310-1322.
12. Schoen RE, Pinsky PF, Weissfeld JL, et al. Colorectal-cancer incidence and mortality with screening flexible sigmoidoscopy. *N Engl J Med.* 2012;366:2345-2357.
13. Fink A, Kosecoff J, Chassin M, Brook RH. Consensus methods: characteristics and guidelines for use. *Am J Public Health.* 1984;74: 979-983.
14. Young GP, Allison J. Screening for colorectal cancer. In: Yamada T, Alpers D, Kaplowitz N, Laine L, Owyang C, Powell D, eds. Textbook of Gastroenterology. 5th ed. Philadelphia, PA: Lippincott Williams and Wilkins; 2008:170-182.
15. Young GP, Cole SR. Which fecal occult blood test is best to screen for colorectal cancer? *Nat Clin Pract Gastroenterol Hepatol.* 2009;6: 140-141.
16. van Rossum LG, van Rijn AF, Laheij RJ, et al. Random comparison of guaiac and immunochemical fecal occult blood tests for colorectal cancer in a screening population. *Gastroenterology.* 2008;135:82-90.
17. Brenner H, Chang-Claude J, Seiler CM, Rickert A, Hoffmeister M. Protection from colorectal cancer after colonoscopy: a population-based, case-control study. *Ann Intern Med.* 2011;154:22-30.
18. Baxter NN, Warren JL, Barrett MJ, Barrett MJ, Stukel TA, Doria-Rose PV. Association between colonoscopy and colorectal cancer mortality in a US cohort according to site of cancer and colonoscopist specialty. *J Clin Oncol.* 2012;30:2664-2669.
19. Nishihara R, Wu K, Lochhead P, et al. Long-term colorectal-cancer incidence and mortality after lower endoscopy. *N Engl J Med.* 2013; 369:1095-1105.
20. Bossuyt PM, Lijmer JG L, Mol BW. Randomised comparisons of medical tests: sometimes valid, not always efficient. *Lancet.* 2000; 356:1844-1847.
21. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ.* 2006; 332:1089-1092.
22. Hiwatashi N, Morimoto T, Fukao A, et al. An evaluation of mass-screening using fecal occult blood-test for colorectal-cancer in Japan—a case control study. *Jpn J Cancer Res* 1993;84:1110-1112.
23. Saito H, Soma Y, Koeda J, et al. Reduction in risk of mortality from colorectal cancer by fecal occult blood screening with immunochemical hemagglutination test. A case-control study. *Int J Cancer.* 1995;61:465-469.
24. Zappa M, Castiglione G, Grazzini G, et al. Effect of faecal occult blood testing on colorectal mortality: results of a population-based case-control study in the district of Florence, Italy. *Int J Cancer.* 1997;73:208-210.
25. Saito H, Soma Y, Nakajima M, et al. A case-control study evaluating occult blood screening for colorectal cancer with hemoccult test and an immunochemical hemagglutination test. *Oncol Rep.* 2000;7:815-819.
26. Nakajima M, Saito H, Soma Y, Sobue T, Tanaka M, Munakata A. Prevention of advanced colorectal cancer by screening using the immunochemical faecal occult blood test: a case-control study. *Brit J Cancer.* 2003;89:23-28.
27. Lee KJ, Inoue M, Otani T, Iwasaki M, Sasazuki S, Tsugane S. Colorectal cancer screening using fecal occult blood test and subsequent risk of colorectal cancer: a prospective cohort study in Japan. *Cancer Detect Prevent.* 2007;31:3-11.
28. Saito H. Screening for colorectal cancer: current status in Japan. *Dis Colon Rectum.* 2000;43(10 suppl):S78-S84.
29. Oort FA, Terhaar Sive Droste JS, Van Der Hulst RW, et al. Colonoscopy-controlled intra-individual comparisons to screen relevant neoplasia: faecal immunochemical test vs guaiac-based faecal occult blood test. *Aliment Pharmacol Ther.* 2010;31:432-439.
30. von Karsa L, Patnick J, Segnan N, et al. European guidelines for quality assurance in colorectal cancer screening and diagnosis: overview and introduction to the full supplement publication. *Endoscopy.* 2013;45:51-59.
31. Halloran SP, Launoy G, Zappa M; International Agency for Research on Cancer. European guidelines for quality assurance in colorectal cancer screening and diagnosis. First edition—faecal occult blood testing. *Endoscopy.* 2012;44(suppl 3):SE65-SE87.
32. Lane JM, Chow E, Young GP, et al. Interval fecal immunochemical testing in a colonoscopic surveillance program speeds detection of colorectal neoplasia. *Gastroenterology.* 2010;139:1918-1926.
33. de Wijkerslooth TR, de Haan MC, Stoop EM, et al. Reasons for participation and nonparticipation in colorectal cancer screening: a randomized trial of colonoscopy and CT colonography. *Am J Gastroenterol.* 2012;107:1777-1783.
34. Cheng H, Macaluso M. Comparison of the accuracy of two tests with a confirmatory procedure limited to positive results. *Epidemiology.* 1997;8:104-106.

35. Cheng H, Macaluso M, Waterbor J. Estimation of relative and absolute test accuracy. *Epidemiology.* 1999;10:566-568.
36. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford, UK: Oxford University Press; 2003.
37. St John DJB, Young GP, Alexeyeff MA, et al. Evaluation of new occult blood-tests for detection of colorectal neoplasia. *Gastroenterology.* 1993;104:1661-1668.
38. Allison JE, Tekawa IS, Ransom LJ, Adrain AL. A comparison of fecal occult-blood tests for colorectal cancer screening. *N Engl J Med.* 1996;334:155-159.
39. Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *BMJ.* 2001;323:157-162.
40. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Evidence base of clinical diagnosis—designing studies to ensure that estimates of test accuracy are transferable. *Brit Med J.* 2002;324:669-671.
41. Young GP, Macrae FA, St John DJB. Clinical methods of early detection: basis, use and evaluation. In: Young GP, Rozen P, Levin B, eds. Prevention and Early Detection of Colorectal Cancer. London, UK: Saunders; 1996:241-270.
42. Cole SR, Young GP, Esterman A, Cadd B, Morcom J. A randomised trial of the impact of new faecal haemoglobin test technologies on population participation in screening for colorectal cancer. *J Med Screen.* 2003;10:117-122.
43. Young GP, St John DJB. Selecting an occult blood test for use as a screening tool for large bowel cancer. *Front Gastrointest Res.* 1991; 18:135-156.
44. Young GP, St John DJ, Winawer SJ, Rosen P, (WHO) World Health Organization and OMED (World Organization for Digestive Endoscopy). Choice of fecal occult blood tests for colorectal cancer screening: recommendations based on performance characteristics in population studies: a (WHO) World Health Organization and OMED (World Organization for Digestive Endoscopy) report. *Am J Gastroenterol.* 2002;97:2499-2507.
45. Young GP. Screening for colorectal cancer: alternative faecal occult blood tests. *Eur J Gastroenterol Hepatol.* 1998;10:205-212.
46. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *CMAJ.* 1986;134:587-594.
47. Smith A, Young GP, Cole SR, Bampton P. Comparison of a brush-sampling fecal immunochemical test for hemoglobin with a sensitive guaiac-based fecal occult blood test in detection of colorectal neoplasia. *Cancer.* 2006;107:2152-2159.
48. Sillars-Hardebol AH, Carvalho B, van Engeland M, Fijneman RJ, Meijer GA. The adenoma hunt in colorectal cancer screening: defining the target. *J Pathol.* 2012;226:1-6.
49. Cole SR, Tucker GR, Osborne JM, et al. Shift to earlier stage at diagnosis as a consequence of the National Bowel Cancer Screening Program. *Med J Aust.* 2013;198:327-330.
50. Levin B, Liberman DA, McFarland, et al; American Cancer Society Colorectal Cancer Advisory Group. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *Gastroenterology.* 2008;134:1570-1595.
51. Atkin WS, Morson BC, Cuzick J. Long-term risk of colorectal cancer after excision of rectosigmoid adenomas. *N Engl J Med.* 1992; 326:658-662.
52. Miles A, Cockburn J, Smith RA, Wardle J. A perspective from countries using organized screening programs. *Cancer.* 2004;101: 1201-1213.
53. Lansdorp-Vogelaar I, Kuntz KM, Knudsen AB, van Ballegooijen M, Zauber AG, Jemal A. Contribution of screening and survival differences to racial disparities in colorectal cancer rates. *Cancer Epidemiol Biomarkers Prev.* 2012;21:728-736.
54. Faivre J, Dancourt V, Lejeune C, et al. Reduction in colorectal cancer mortality by fecal occult blood screening in a French controlled study. *Gastroenterology.* 2004;126:1674-1680.
55. Cole SR, Gregory T, Whibley A, et al. Predictors of re-participation in faecal occult blood test-based screening for colorectal cancer. *Asian Pac J Cancer Prev.* 2012;13:5989-5994.
56. Chen LS, Liao CS, Chang SH, Lai HC, Chen TH. Cost-effectiveness analysis for determining optimal cut-off of immunochemical faecal occult blood tests for population-based colorectal cancer screening. *J Med Screen.* 2007;14:191-199.
57. Ransohoff DF. Rules of evidence for cancer molecular marker discovery and validation. *Nat Rev Cancer.* 2004;4:309-311.
58. Imperiale TF Ransohoff DF, Itzkowitz SH, et al. Multitarget stool DNA testing for colorectal-cancer screening. *N Engl J Med.* 2014; 370:1287-1297.